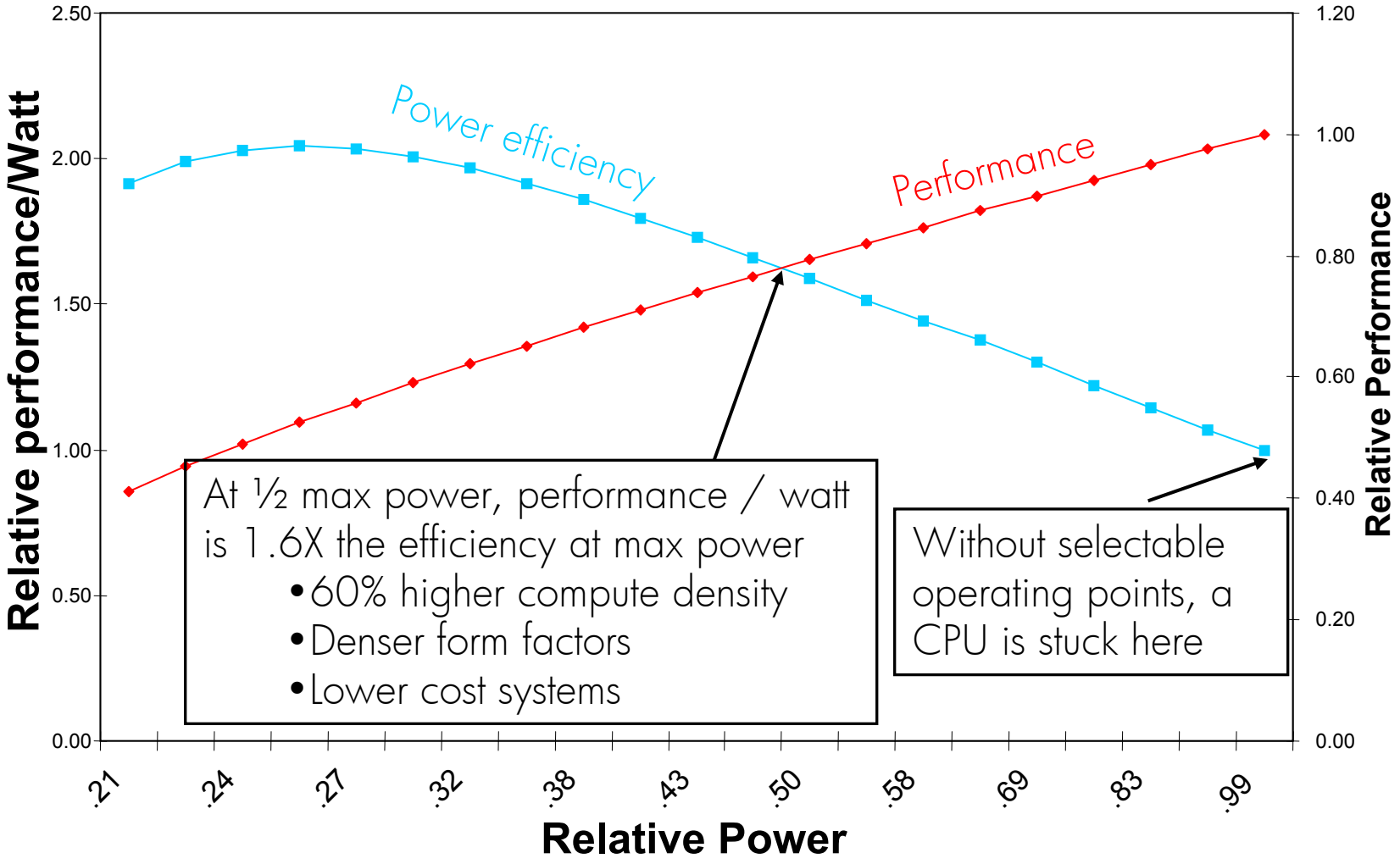
Power Management in Future IPF Processors

Sam Naffziger
Hewlett Packard Co., Fort Collins, CO

Motivation – Quotes from Fall '02 Press

- "AMD is hoping that Itanium 2 will be hampered by a higher price tag, higher power consumption, and the inability to run 32-bit x86 software at native speeds"
- It turns out, Schmidt (Google CEO) told the audience that what matters most to the computer designers at Google is **not speed, but power – low power** … if power efficiency does indeed trump processing speed everything Intel and HP have done to pack raw power into [McKinley] could now be a handicap
- The market loves 1U and 2U rack servers, which pack the greatest number of systems into a square foot, affordably. ... **such slim boxes have less room for the cooling devices required by Intel's latest and hottest-running chips**, **Itanium** and Xeon. Dell says it expects to soup up the 1U and 2U boxes' fans and cooling packs, **but the most powerful servers still require the larger form factor.**"
- Regarding entering the rapidly expanding *bladed server* space: "It's difficult to squeeze processors into small spaces without overheating the computer, a problem that can cause data corruption and crashes. Putting higher-end processors such as Xeon and Itanium into small cabinets is trickier than cooler chips such as Intel's ultralow-voltage Pentium III chips"

Motivation: Power/Performance

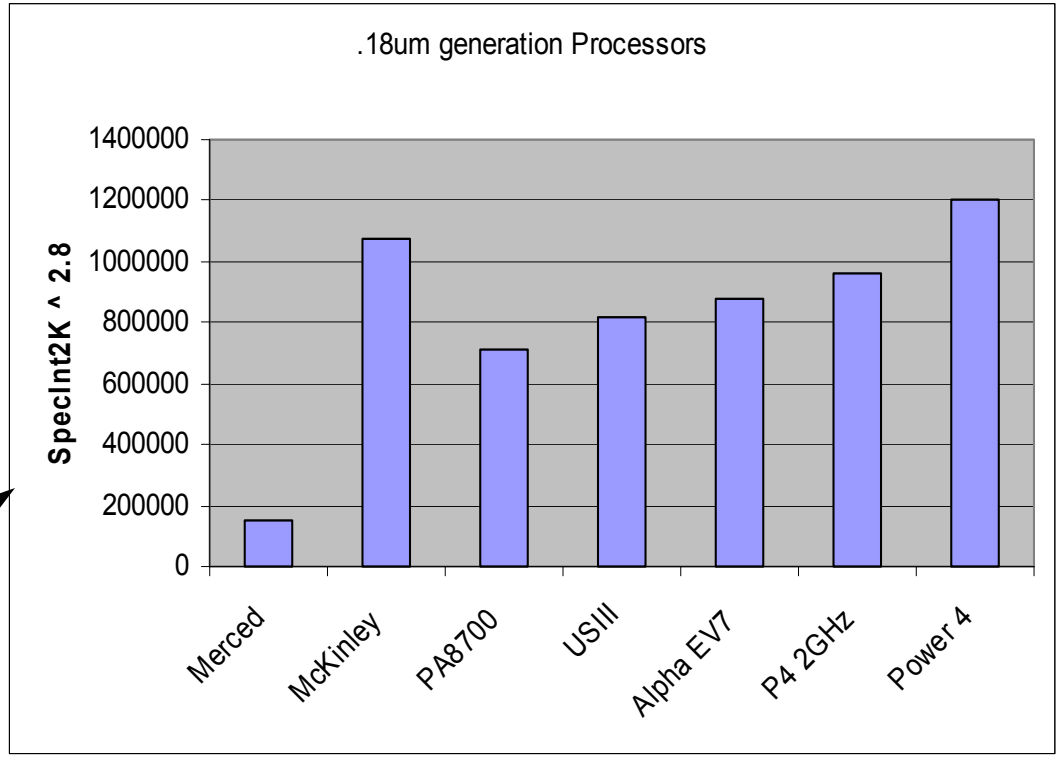


Energy Consumption in Microprocessors

Power $\propto C \cdot V^2 \cdot F$, $F \propto V$, performance $\propto F \Rightarrow$ **Power $\propto$ (perf)^3**
This cubic is diluted by leakage and I/O power $\Rightarrow$ 2.8 is a good measure in 90nm
[Reduce Voltage and Frequency to hit power parity, then measure performance]

For .18µm designs, power efficiency is similar across architectures.

Processor	SpecInt 2K	Watts	Spec^2.8/ Watt
Merced	404	130	152730
McKinley	810	130	1071062
PA8700	569	73	709564
USIII	605	75	820075
Alpha EV7	804	155	879803
P4 (2GHz)	640	75	959937
Power 4	808	115	1202414



Power Management in the Data Center vs. Mobile

- Mobile power management tends to focus on *minimizing energy usage*
  - Conserve batteries by only burning power when needed
  - Performance on demand only
  - Changes not user initiated
- Data Center power management needs to focus on sustained *energy efficiency* and thermals
  - Sustaining maximum performance within the minimum power envelope to maximize computes / ft^3
  - Enable power based data center management
    - Based on utilization or center-level cooling issues
    - Power limits per rack, per center etc.
  - A primary issue is heat removal and associated form factor issues

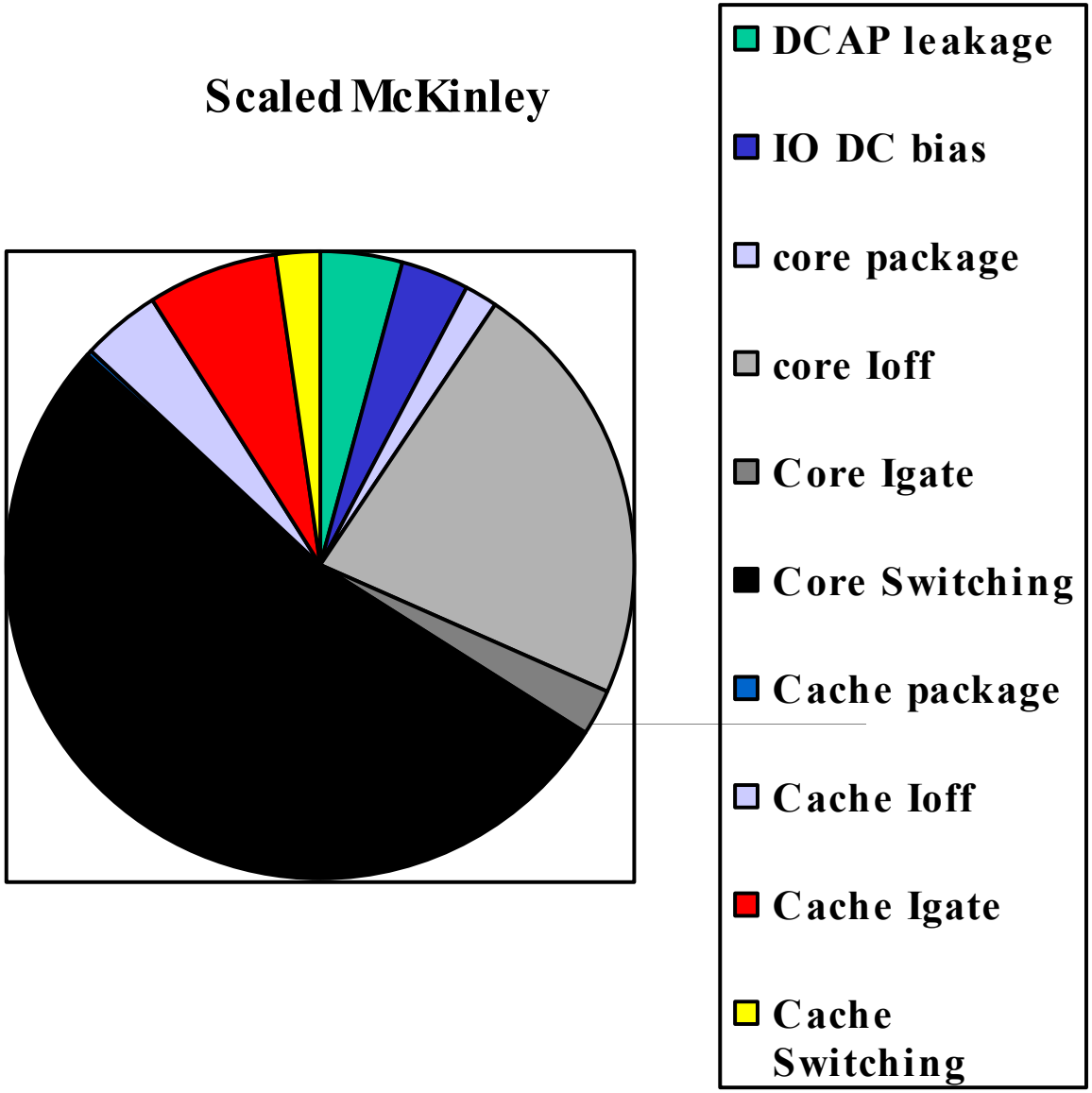
Current Approaches to Power Management and Reduction

Guardband removal

- Use "Thermal Design Power" (TDP) to spec a sustained power that is lower than the true maximum
  - Counts on the rarity of very high activity factor applications and the priority of thermal issues over power delivery
  - Splits out thermal and power delivery issues
- Dynamic Voltage Scaling
  - 2 or more voltage/frequency pairs with an operating system selecting the mode for energy conservation
    - Sometimes mode based (battery/AC power), sometimes activity based (processor utilization high → higher frequency)
- Fuse in a Vcc that is part-specific
  - Higher power but faster parts can use a lower voltage

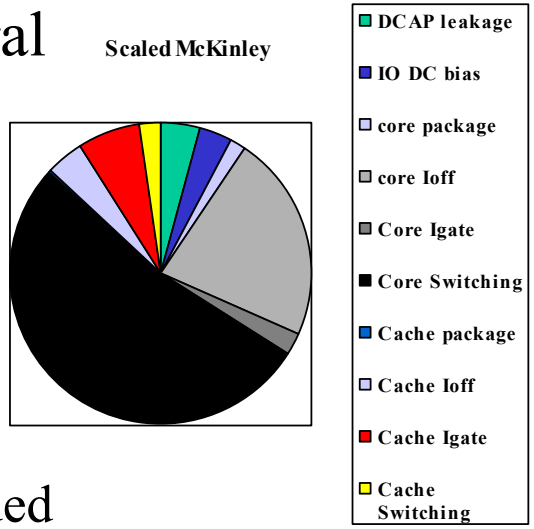
Power Consumption Components

- Taking the 130W McKinley design, porting to 90nm as a dual core with larger caches
- The power consumption is about 3X over budget

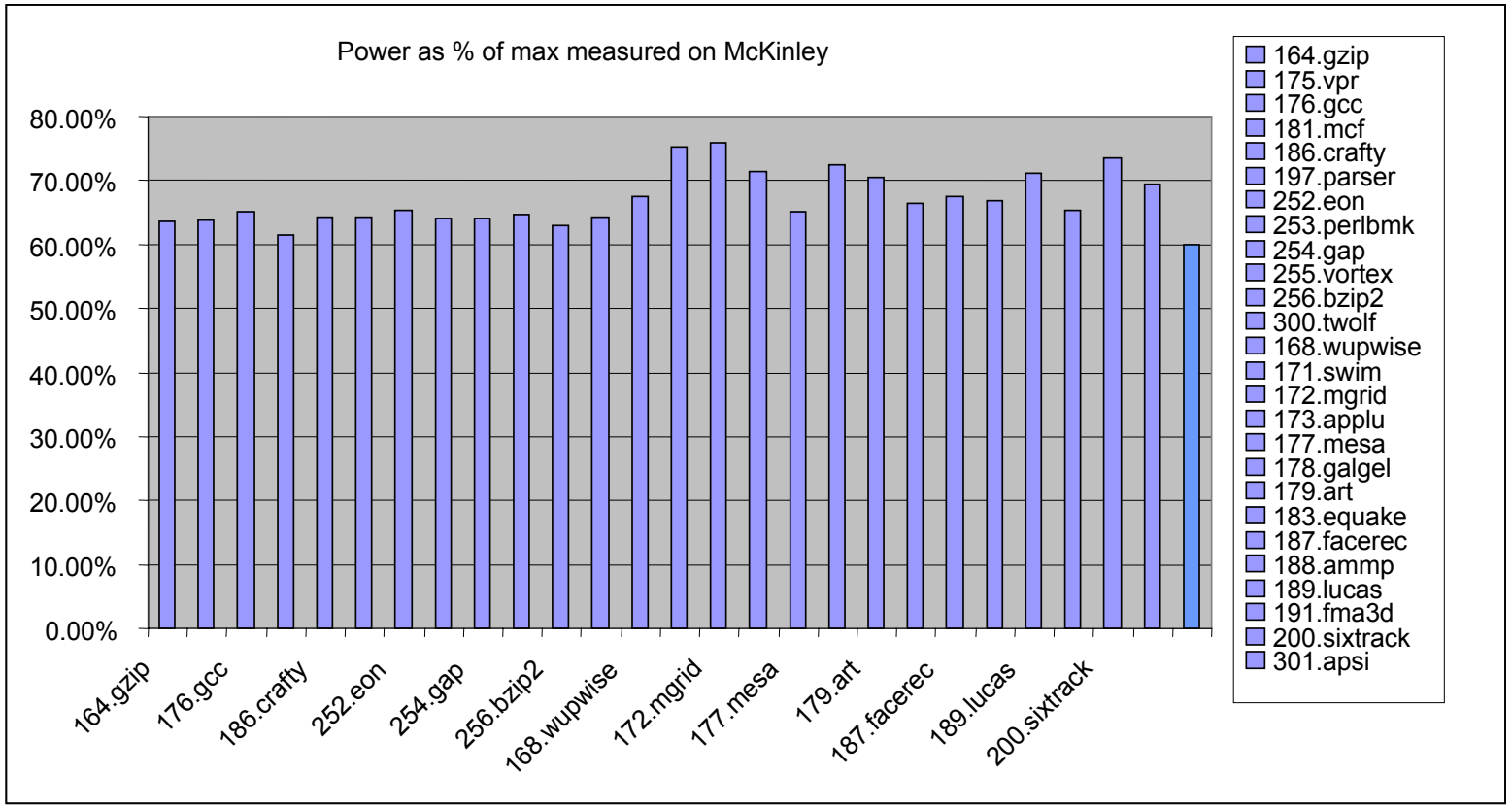


Power Reduction Needs

- To get power to an acceptable level, several approaches must be taken
  - Eliminate as much guardbanding as possible
    - Application power vs. max power
    - Process variability
    - Temperature uncertainty
    - Voltage variability
  - Only apply a high voltage where speed is needed
    - Cache supply can be independent
    - DCAP placement must be carefully scrutinized
  - Reduce the base power consumption of the design
    - Switching cap through clock gating and device size optimization
    - Leakage through sizing and a number of circuit level techniques



Application Power vs. Max Power



- Average application power is in the range of 60% to 75% of max

Optimizing for *typical power* in a highly parallel machine could have a huge return

Application Power vs. Max Power

- The Single biggest opportunity for power reduction
- Two possibilities for converging max power down to application power
  - Use thermal monitoring and spec a TDP
    - Only helps the heat removal, not power delivery
    - Thermals are an indirect indication of power consumption
  - Actually measure power and determine throttling needs based on V*I
    - Average code would run at full frequency
    - Higher activity factor code could have its frequency reduced
    - Benefits both thermals and power delivery as well as power budgets for boxes, racks and datacenters

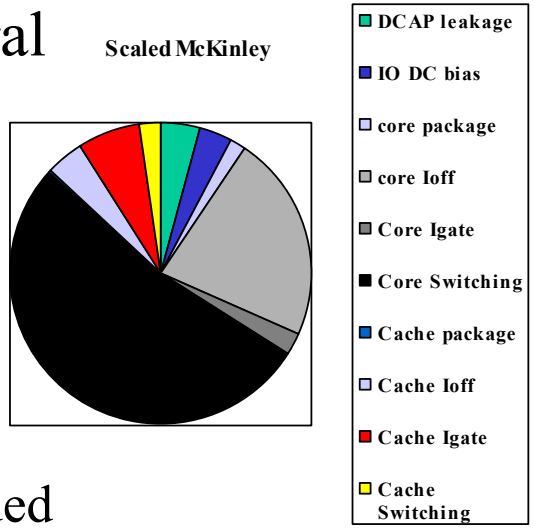
Application Power vs. Max Power

- To implement this approach, we need
  - An accurate method of measuring power
    - Voltage is easy, amperage is hard
    - Inaccuracy must be eaten as lost power
  - A *graceful* means of reducing frequency for applications that exceed the power limit
    - If (frequency redux/power redux) is too high, the throttle threshold will have to be set high to avoid dinging high AF* code too much
    - If V/F pairing is too coarse and switch time long between pairs, code could perform unpredictably and be victimized by a previous high AF process

* AF = Activity Factor, defined here as % of "power virus" switching

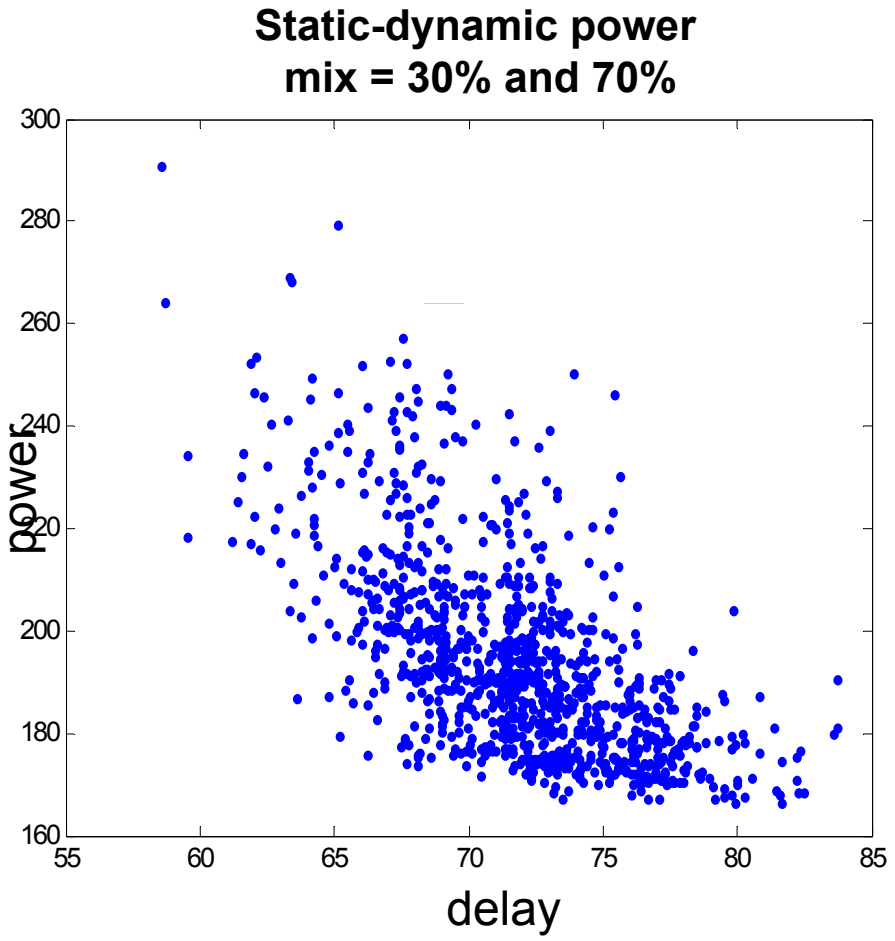
Power Reduction Needs

- To get power to an acceptable level, several approaches must be taken
  - Eliminate as much guardbanding as possible
    - Application power vs. max power
    - Process variability
    - Temperature uncertainty
    - Voltage variability
  - Only apply a high voltage where speed is needed
    - Cache supply can be independent
    - DCAP placement must be carefully scrutinized
  - Reduce the base power consumption of the design
    - Switching cap through clock gating and device size optimization
    - Leakage through sizing and a number of circuit level techniques



Reducing Process Variation Impacts

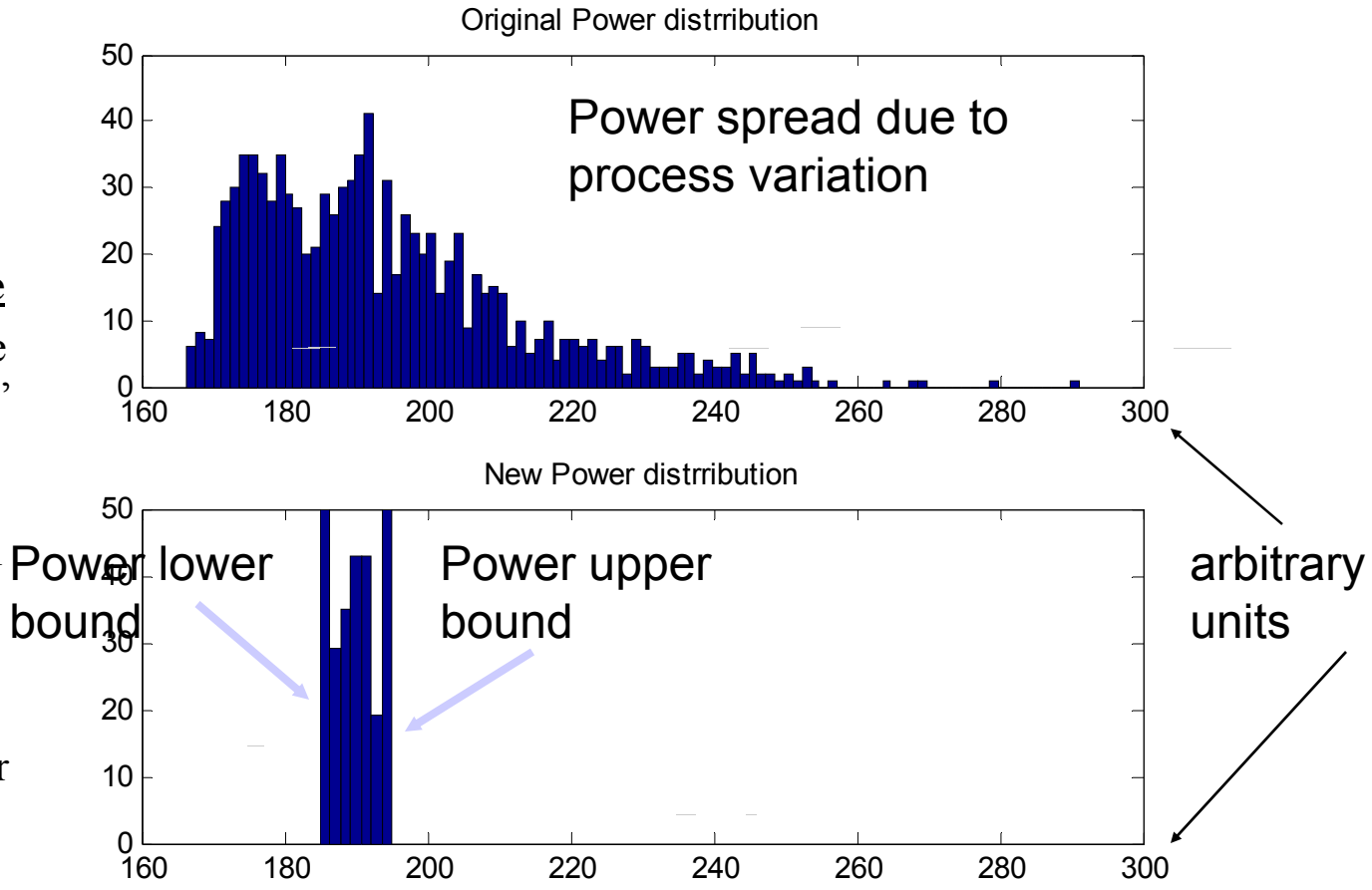
- Leakage is now 30%-50% of the total power budget
- It is also one of the least controlled and variable of process parameters
- High leakage tends to correlate well with faster devices



Adjusting the Power To the Nominal Power by the Upper and Lower Bounds

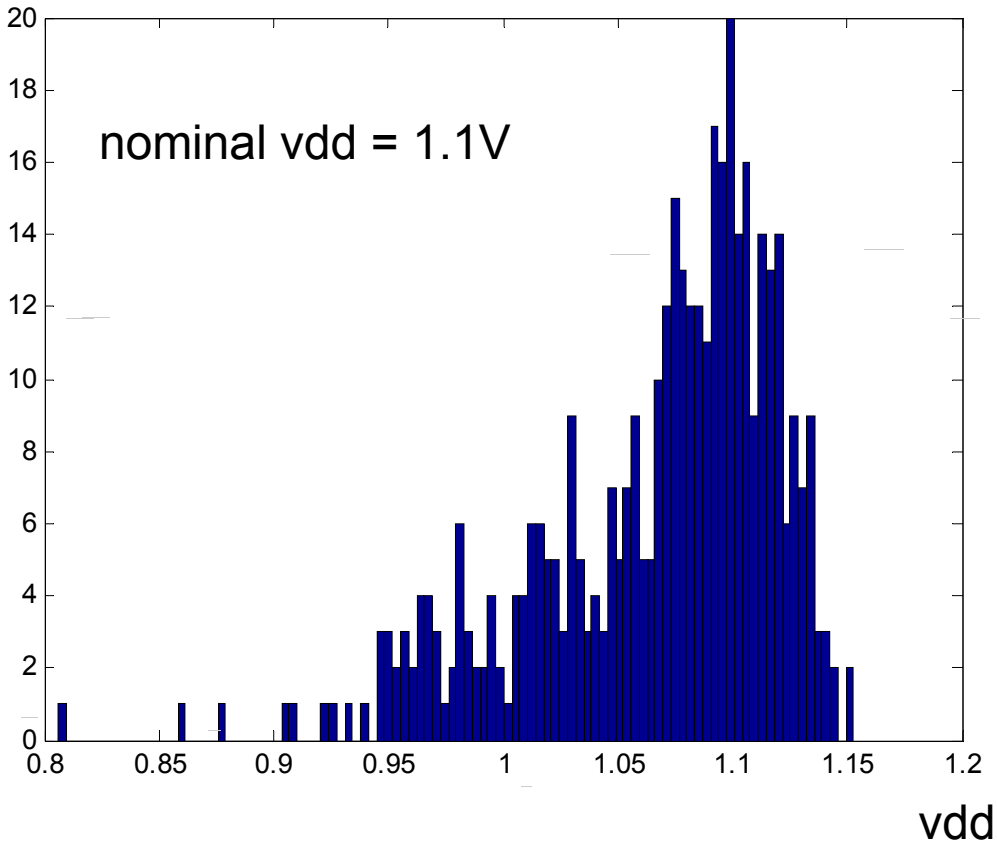
- IF we can **measure power** and **manage it gracefully**
- AND **manage our own voltage**
- We can bring the high power "tail" of parts coming out of the fab down to nominal power

➢ Eliminate power envelope guardbanding for those parts



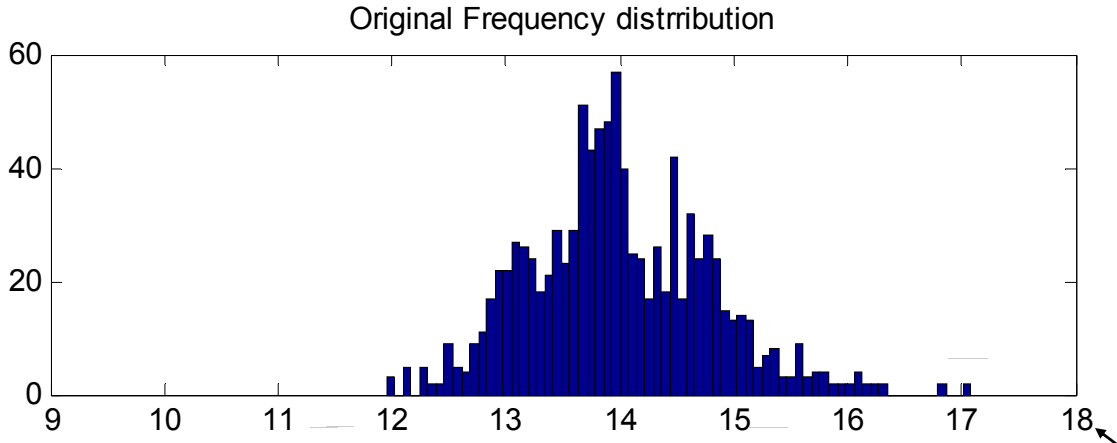
Vdd Dist. Bounded At the Nominal Power Point (histogram)

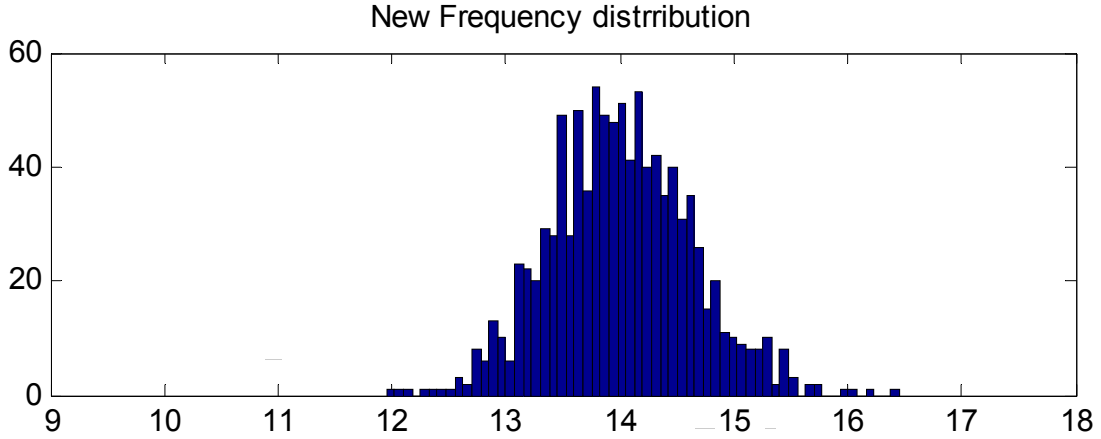
- Each part ends up with a unique, optimal operating voltage at the power limit
- Fine grained voltage control makes this more effective
- Only useful for a *power limited* design – otherwise all parts would operate at maximum Vdd
- But, what happens to the frequency?



Experimental Results: Modified Freq. Distribution (Nominal Power)

- Some tightening of the distribution
- Nominal bin split unaffected by a 20-25% reduction in power

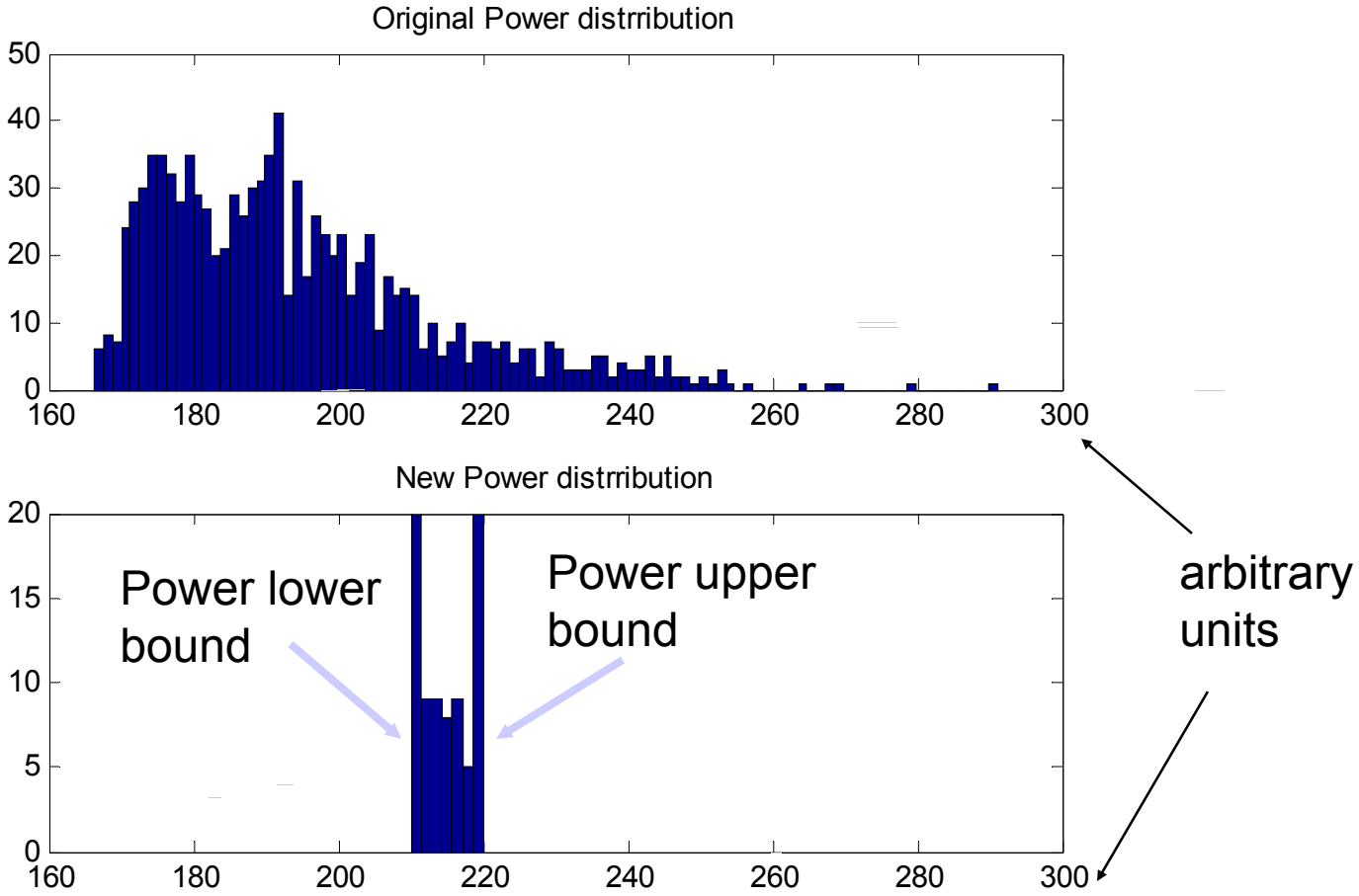




arbitrary units

Static-dynamic power mix = 30% and 70%

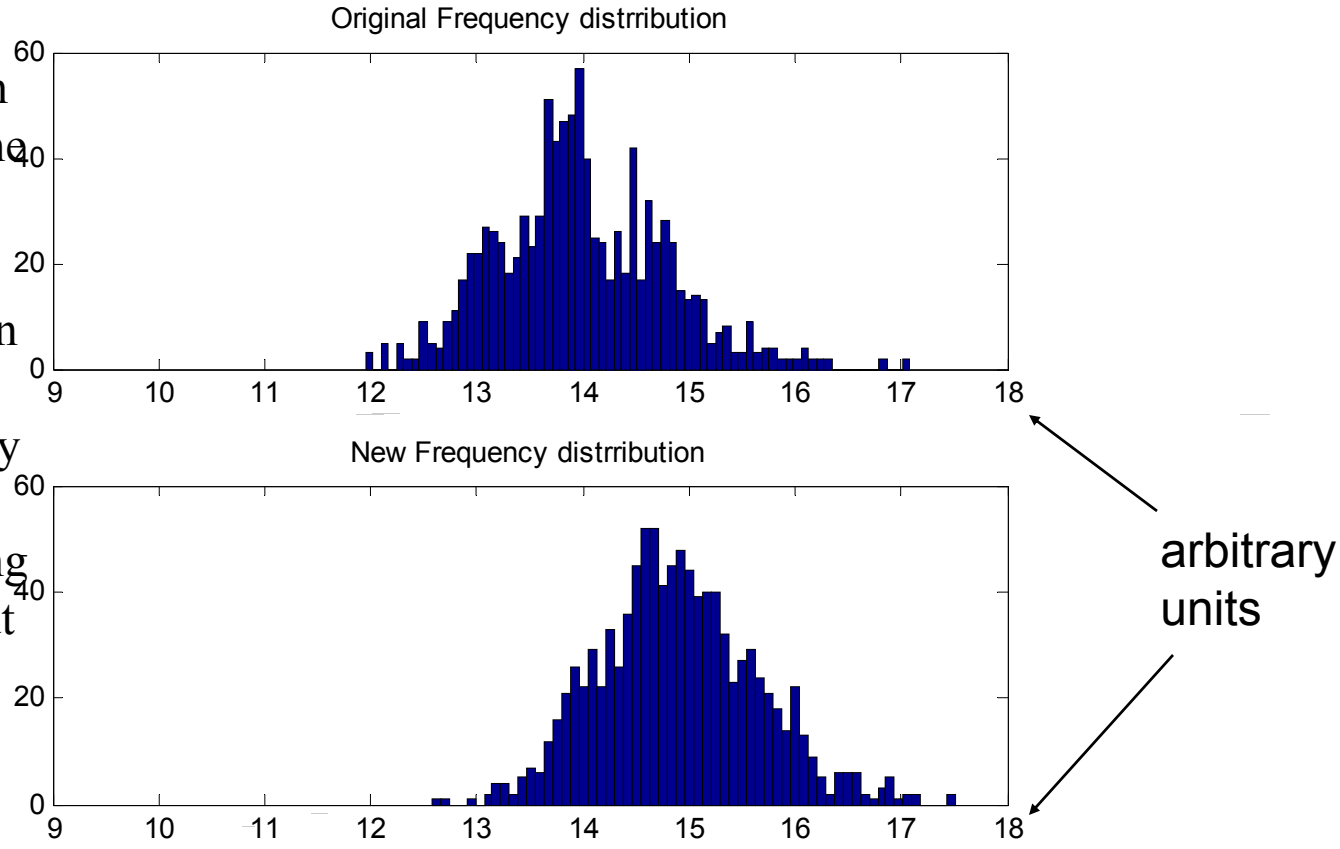
Another option: Adjusting the Power to 90% Max



Static-dynamic power mix = 30% and 70%

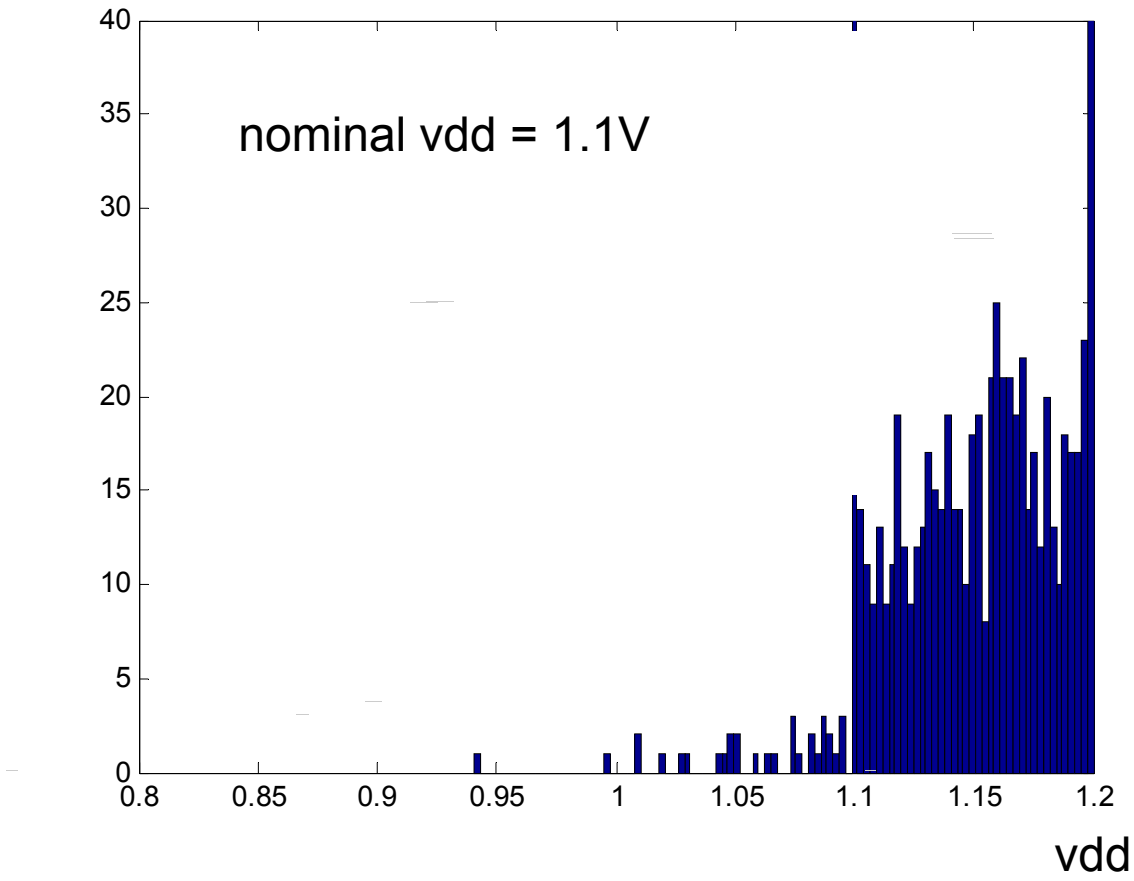
Experimental Results: Modified Freq. Distribution at 90% Max Power)

- Distribution moved to the right about 7%
- Nominal bin split improved by 7% while still reducing power about 10%



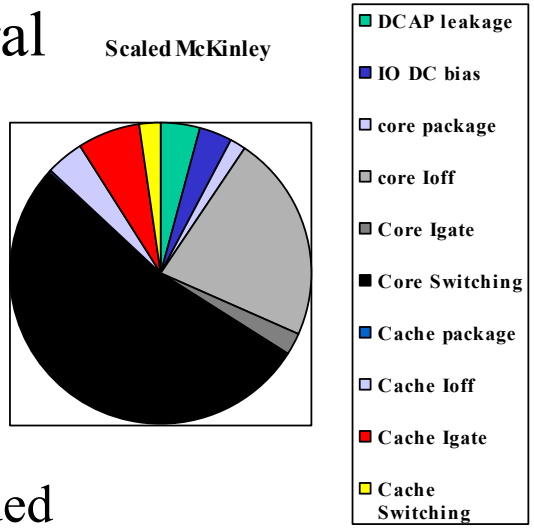
Static-dynamic power mix = 30% and 70%

Vdd Dist. Under QR Bounded at the 90% Power



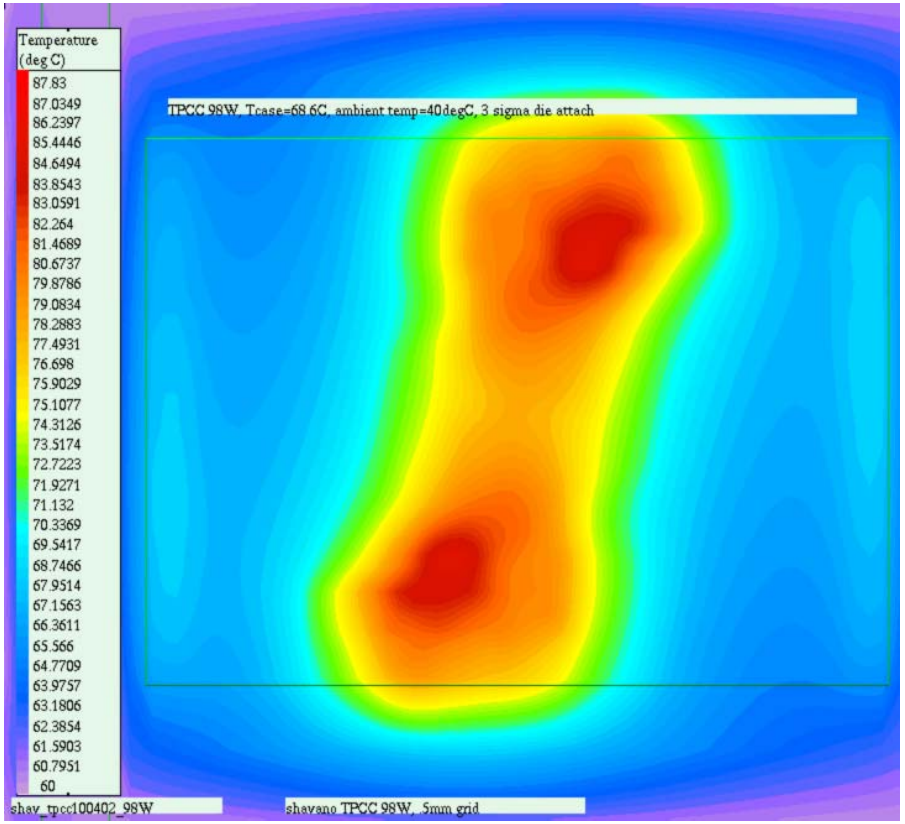
Power Reduction Needs

- To get power to an acceptable level, several approaches must be taken
  - Eliminate as much guardbanding as possible
    - Application power vs. max power
    - Process variability
    - Temperature uncertainty
    - Voltage variability
  - Only apply a high voltage where speed is needed
    - Cache supply can be independent
    - DCAP placement must be carefully scrutinized
  - Reduce the base power consumption of the design
    - Switching cap through clock gating and device size optimization
    - Leakage through sizing and a number of circuit level techniques



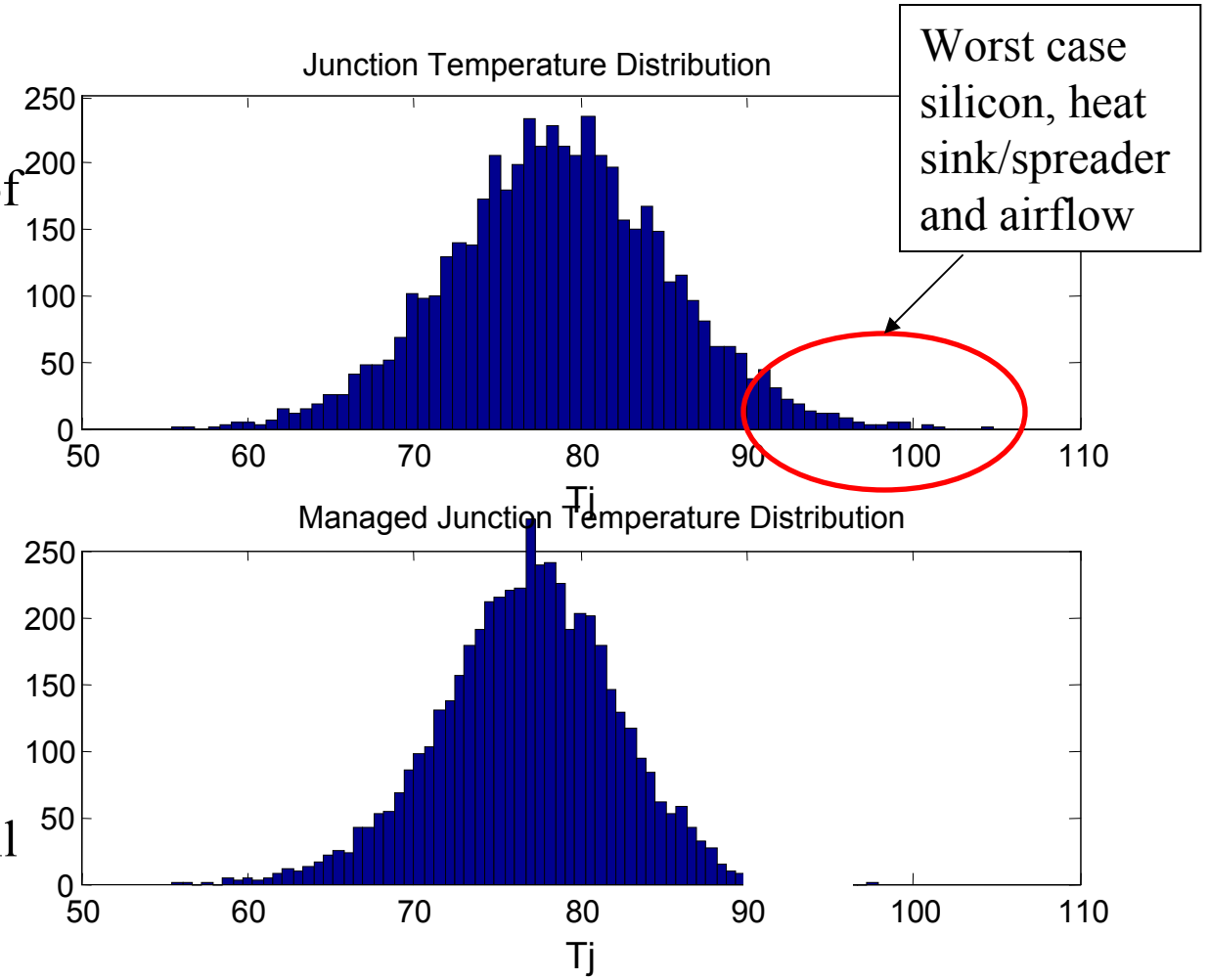
Temperature Management

- Higher junction temperature results in slower transistors, higher interconnect resistance and exponentially higher subthreshold leakage
- Reduced power consumption results in reduced Tj ➔ a virtuous cycle



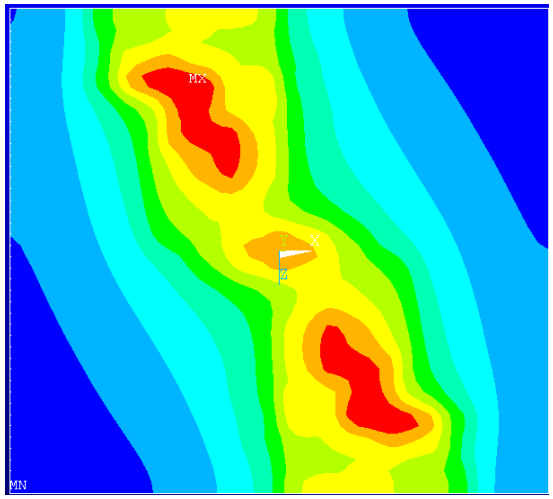
Temperature Management

- IF we have an accurate means of temperature monitoring
- AND a *graceful* throttle mechanism
- We can avoid making all parts pay in performance for the worst case tail



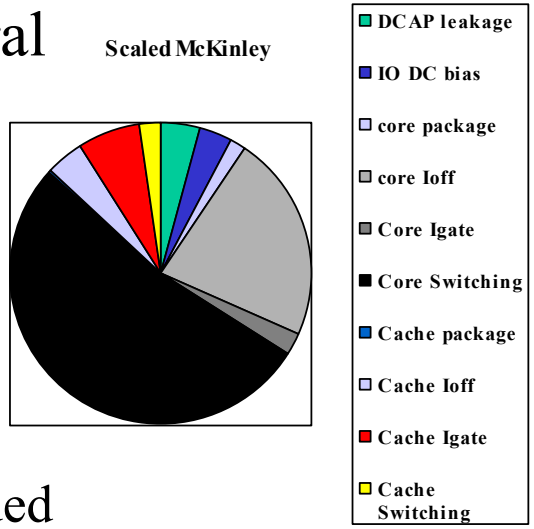
Temperature Management

- By improving the accuracy of temperature monitoring and incorporating in our power control loop, we gain:
  - The ability to reduce temperature guardband due to variations in parts and operating conditions
  - A resulting reduction in leakage power and increase in silicon performance
  - And an increase in RAS
- The performance impact can be kept below a noticeable threshold for the occasional occurrence by properly setting the limit



Power Reduction Needs

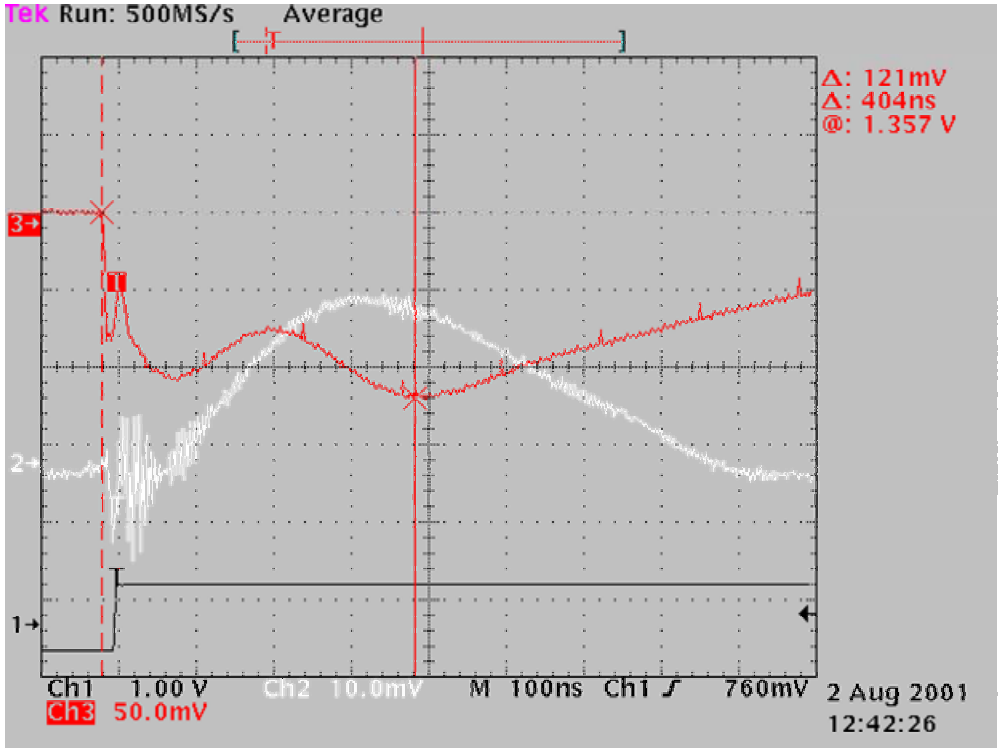
- To get power to an acceptable level, several approaches must be taken
  - Eliminate as much guardbanding as possible
    - Application power vs. max power
    - Process variability
    - Temperature uncertainty
    - Voltage variability
  - Only apply a high voltage where speed is needed
    - Cache supply can be independent
    - DCAP placement must be carefully scrutinized
  - Reduce the base power consumption of the design
    - Switching cap through clock gating and device size optimization
    - Leakage through sizing and a number of circuit level techniques



The Impact of Voltage Variation

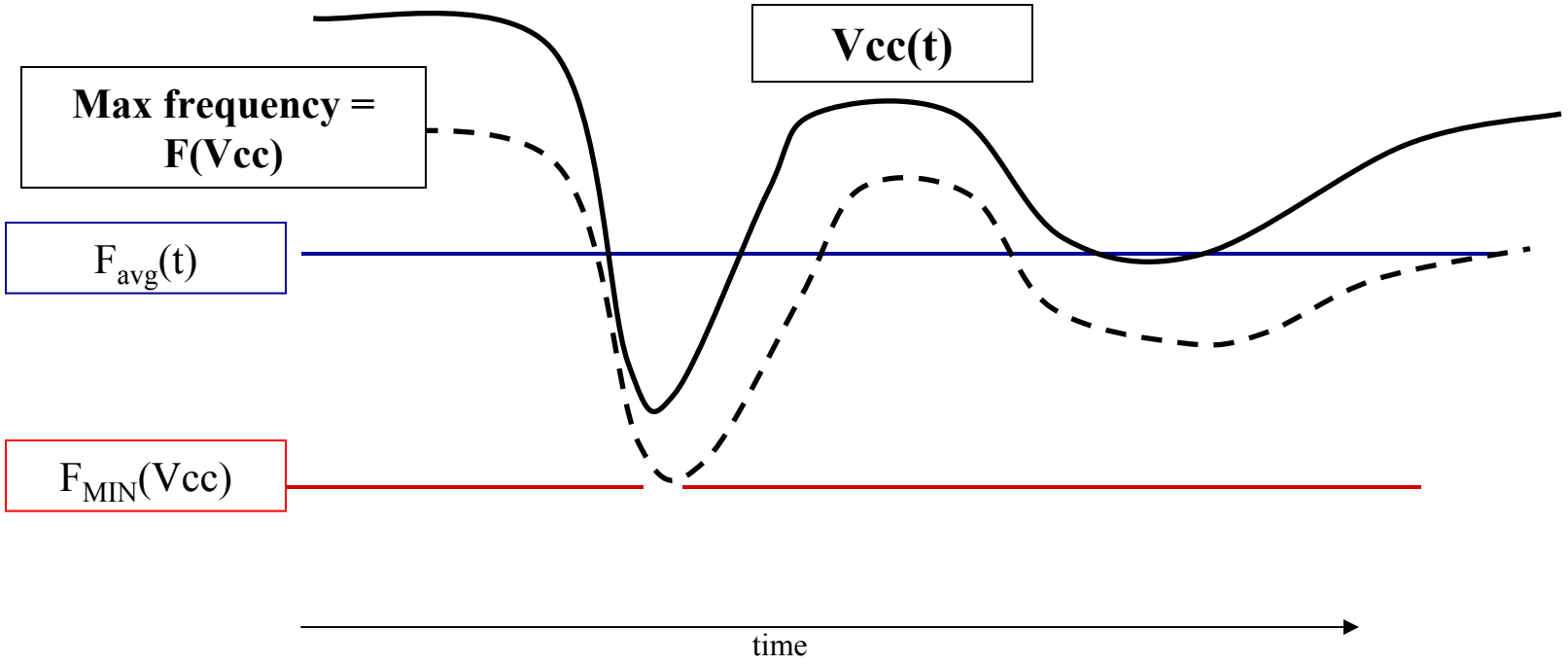
- Sudden changes in power consumption induce voltage transients
- ➢ To avoid a timing problem, average voltage must be higher than needed except for the rare droop case ➔ $[(V_{avg}+V_{droop})/V_{avg}]^2$

Measured McKinley: 143mV droop With a 32A step load (~33%) At 1.5V



The Impact of Voltage Variation

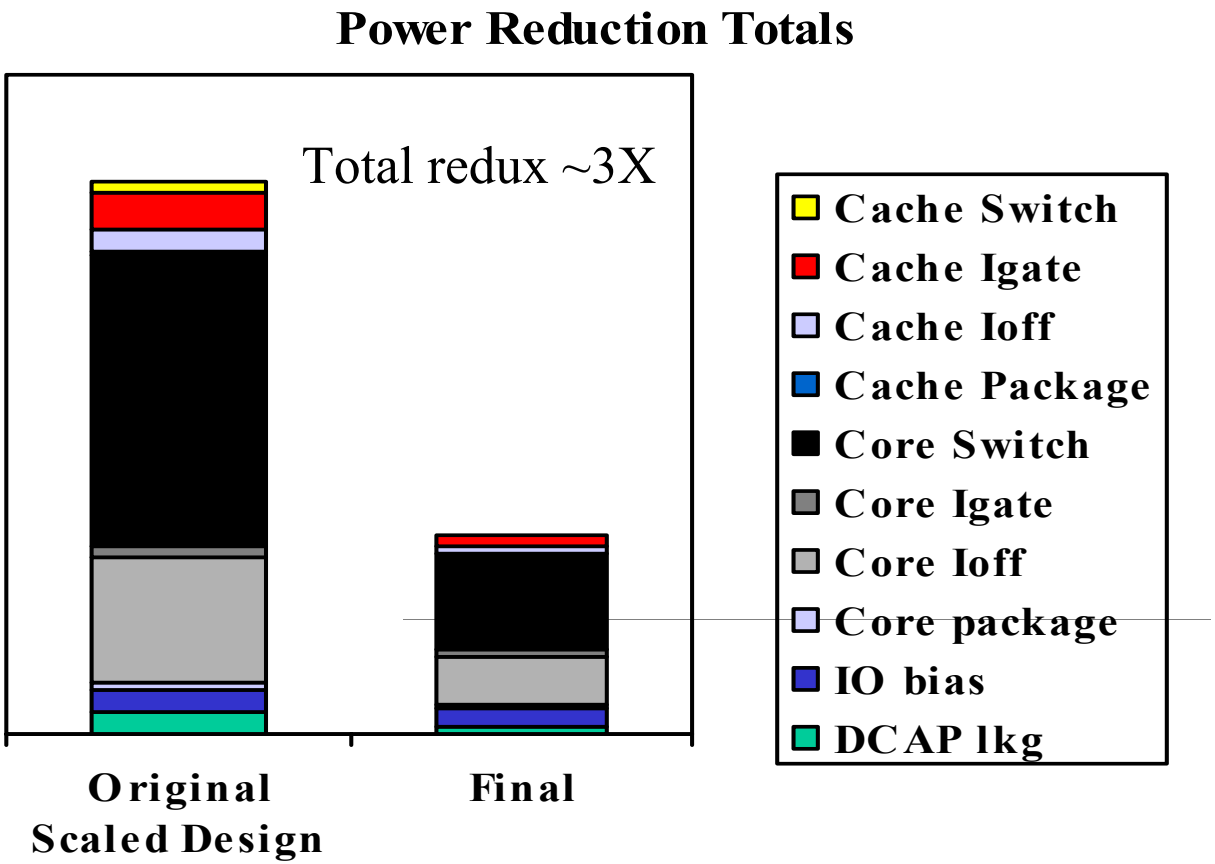
- Traditional designs must operate at $F_{MIN}(Vcc)$ to avoid a speedpath failure
- IF we could design a clock generation system that responds rapidly to voltage transients, we can ride the frequency at an $F_{AVG}(t)$ that is much closer to the maximum allowable by the average, not minimum Vcc
  - $[V_{avg}/V_{min}]^2$
  - If $V_{avg} = 1.1X\ V_{min}$, the power savings is > 20%



Some More Basic Approaches as Well

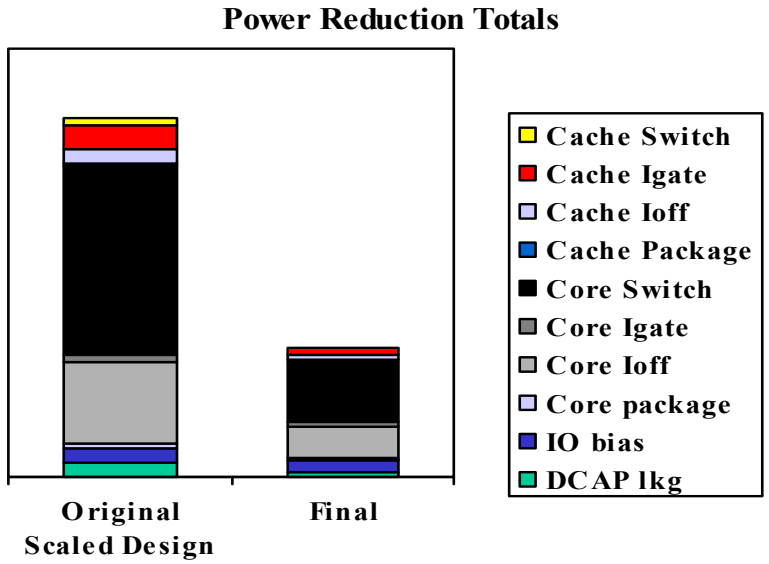
- Separate out the supply for the large highest level cache which is latency insensitive
  - Leakage reduced exponentially with voltage
- Implement extensive clock gating and logic qualification to reduce switching cap
- Reduce total FET width and use non-minimum length devices for leakage reduction
- Eliminate as much domino and contention based logic as possible

Power Reduction Breakdown



Power Reduction Summary

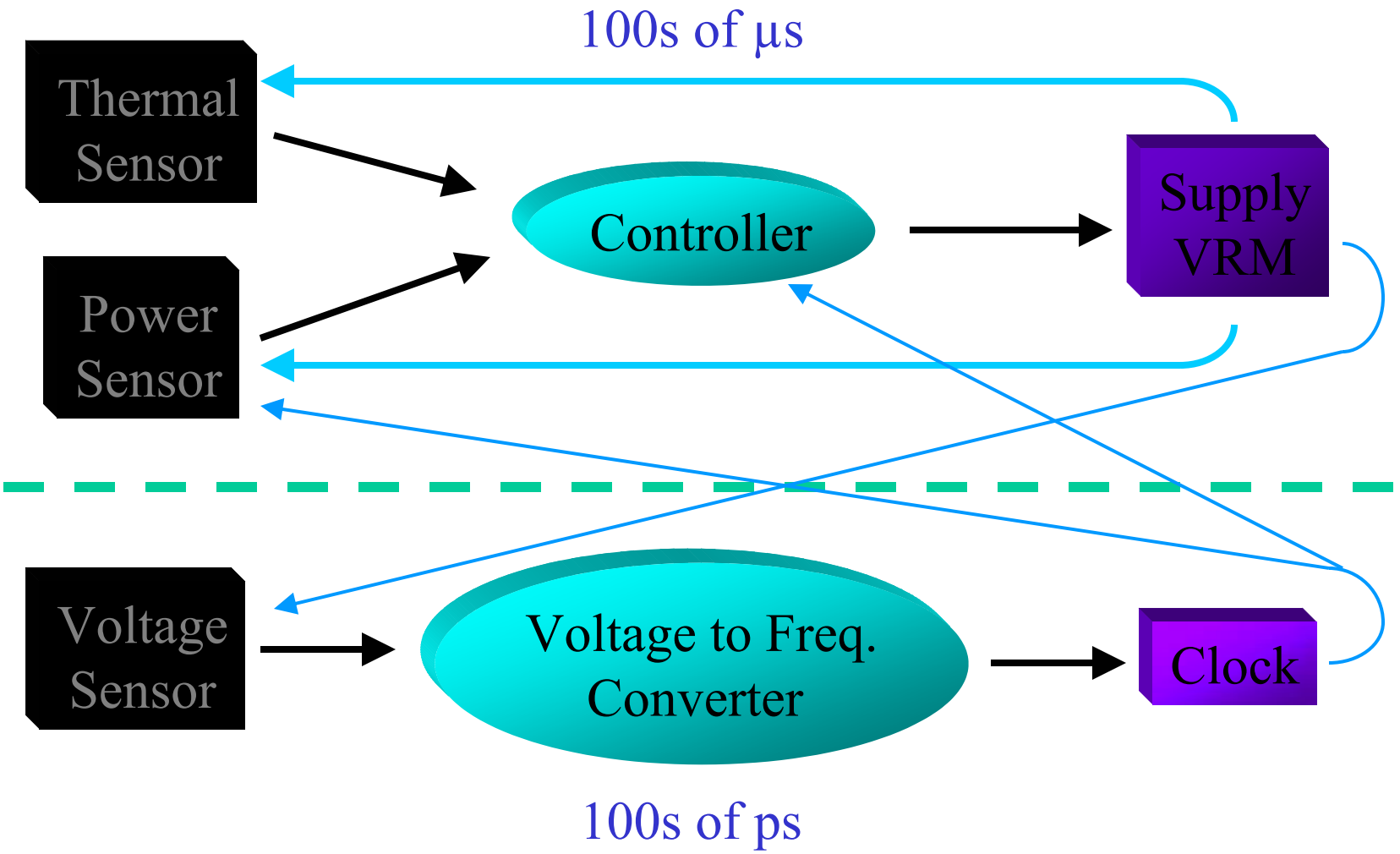
1) Manage to application power vs. max power
2) Adapt Vcc to optimal value for each part
3) Manage junction temperature to the minimum possible
4) Adapt frequency to Vcc changes to operate at frequency(V_{avg}) vs. frequency(V_{min})
5) Separate supply for cache
6) Optimize circuits for low switching and low leakage



Key enablers:

1) High accuracy ammeter
2) Dynamic voltage control
3) Fast frequency synthesis as a function of Vcc
4) Accurate thermal measurement

10,000 Foot View of System



Stuff I Didn't Cover

- How do you market dynamic frequency variation?
- How to you test and verify a self-clocking, environmentally aware part?
- Can a frequency synthesizer really respond sufficiently fast to voltage changes?
- Can a sufficiently accurate ammeter be designed without burning too much power of its own?
- If the part is power and thermal aware, can its operating power be set dynamically?

Power Control Simulation with Voltage Modulation to Adapt to Software Work-load